A FRAMEWORK FOR OFFLINE HANDWRITTEN DEVANAGARI WORD RECOGNITION

A

RESEARCH PROPOSAL

SUBMITTED

IN PARTIAL FULFILLMENT TO THE REQUIREMENT

FOR

DOCTOR OF PHILOSOPHY

IN

FACULTY OF ENGINEERING & TECHNOLOGY

ELECTRONICS & COMMUNICATION ENGINEERING

By

Sukhjinder Singh
(Enrolment No. 1610AMPE01)

MAHARAJA RANJIT SINGH PUNJAB TECHNICAL UNIVERSITY

BATHINDA (INDIA)

(August, 2018)

(Final R Proposal Recommended by DDRC as on 14.08.2018)

TABLE OF CONTENTS

1.1 INTRODUCTION ……...… 1

1.2 HANDWRITTEN WORD RECOGNITION …….. 1 - 3

1.2.1 Online Handwritten Word Recognition …….. 2

1.2.2 Offline Handwritten Word Recognition ……..... 2 - 3

1.2.2.1 Segmentation Based or Analytical Approach ……..……. 3

1.2.2.2 Segmentation Free or Holistic Approach …….……. 3

1.2.2.3 Hybrid ……...……... 3

1.3 APPLICATIONS AND CHALLENGES ……..... 3 - 5

1.3.1 Applications ……...…. 4 - 5

1.3.2 Challenges ……... 5

1.4 OVERVIEW OF DEVANAGRI SCRIPT ……..……. 6 - 7

2.1 LITERATURE REVIEW ……...……...….. 7

2.2 RELATED WORK …….. 7 - 13

2.2.1 Numeral/Digit Recognition……..…...……... 7 - 8

2.2.2 Word/Text Recognition ……...……...… 8 - 9

2.2.3 Isolated Character Recognition ……...……...…. 9 - 11

2.2.4 Script Recognition ……...…… 11 - 12

2.2.5 Applications ……...…...…….. 12 - 13

3.1 PHASES OF OCR SYSTEM …….……...……... 13 - 15

4.1 RESEARCH ISSUES/GAPS ……...…… 15 - 16

4.2 PROBLEM FORMULATION……..……... 16 - 17

4.3 CONTRIBUTION AND NEED OF THE PROPOSED RESEARCH …… 17 - 18

5.1 RESEARCH OBJECTIVES …… 18

5.2 AN OUTLINE OF THE RESEARCH PROPOSAL …… 19 - 20

5.2.1 Methodology ……... 19

5.2.1.1 Assumptions…….. 19

5.2.2 Plan of Work ……...…… 20

BIBLIOGRAPHY ……..…….. 21 - 25

****

LIST OF FIGURES

Figure No.	Caption	Page No.
1.1	Streams of handwritten word recognition ……..	2
1.2	Devanagari script: (a) Consonants and corresponding half forms (b) Vowels (c) Modifiers……...	6
1.3	Three strips of a word in Devanagari script ……	7
3.1	Phases of OCR systems ……..……	13

****

LIST OF TABLES

Table No.	Caption	Page No.
1.1	Offline *Vs* online handwritten word recognition ……..……..	2
5.1	Plan of work (Time-line chart)…….....	20

****

LIST OF ABBREVIATIONS

ANN :	Artificial Neural Network
ASCII :	American Standard Code for Information Interchange
BayesNet :	Bayesian Network
CM :	Chebyshev Moments
CNN :	Convolutional Neural Network
DiscCNN :	Discriminative Convolutional Neural Network
ELM :	Extreme Learning Machine
FURIA :	Fuzzy Unordered Rule Induction Algorithm
GA :	Genetic Algorithm
GLAC :	Gradient Local Auto-Correlation
HCR :	Handwritten Character Recognition
HMM :	Hidden Markov Model
HWR :	Handwritten Word Recognition
IPA :	Indian Postal Automation
ISI :	Indian Statistical Institute
ISIDCHAR :	ISI Devanagari Characters *(Database)*
k-NN :	k-Nearest Neighbors
LibLINEAR :	Library for Linear classifier
MICR :	Magnetic Ink Character Recognition
MLP :	Multi-Layer Perceptron
MLP-NN :	Multi-Layer Perceptron-Neural Network
MNIST :	Modified National Institute of Standards and Technology *(Database)*
MQDF :	Modified Quadratic Discriminant Function
NN :	Neural Network
NSHP-HMM :	Non-Symmetric Half Plane-Hidden Markov Model
OCR :	Optical Character Recognition
OSI :	Offline Script Identification
PNN :	Probabilistic Neural Network
RT :	Radon Transform
SCF :	Statistical and Contour based Features
SE :	Spatial Energy
SIW-13 :	Script Identification in the Wild 13 Classes *(Database)*
SL :	Simple Logistic
SVM :	Support Vector Machine
V2DMDCHAR :	Vikas J Dongre and Vijay H Mankar's Characters *(Database)*
WE :	Wavelet Energy

****

1.1 INTRODUCTION

Owing to technology, machine and human interface is very imperative in today's era. This interface results in pattern recognition, artificial intelligence, cognitive computing, search engine and social media more user-friendly and interactive. Optical Character Recognition (OCR) is the part of pattern recognition that attempts to build and design a computer system that can convert scanned images of machine printed or handwritten text (numerals, letters and symbols), into corresponding ASCII (American Standard Code for Information Interchange) characters or machine-readable digital form, automatically. It is generally used to improve the speed of operation, reduce errors or noise in the documents and decrease storage space needed for paper documents.

OCR systems viz. Handwritten Character Recognition (HCR) and Handwritten Word Recognition (HWR) systems save human energy and cost by automatically translating document images into equivalent codes. Most OCR systems are script specific in the sense that they can read characters written in one particular script only (Ghosh *et al.*, 2010). In the Indian subcontinent, a number of languages are in use and an automatic recognition of printed and handwritten scripts facilitates number of applications such as image document sorting and penetrating online libraries of image documents (Sharma and Dhaka, 2016) In India, 23 different languages (including English) and 13 different scripts (including Roman) exist (Obaidullah *et al.*, 2017b). In general, OCRs are script specific, and processing documents having more than one script is not easy.

1.2 HANDWRITTEN WORD RECOGNITION

Handwriting recognition has been the topic of research for many decades (Liwicki *et al.*, 2014). It is a process in which hand-printed (graphical) expressions of a language are transformed into their symbolic representation (Kumar, 2009). Nowadays, one interesting, complex and challenging field in OCR is Handwritten Word Recognition (HWR) due to various writing styles of different individuals. As the handwritten word is mixture of cursive and non-cursive parts so word recognition is significantly difficult. The accuracy of HWR system highly depends upon the feature extraction and classification methods considered. Many researchers had proposed variety of feature extraction and classification methods for

various scripts including Devanagari. Handwritten word recognition has two streams (Dasgupta *et al.*, (2016); Madhvanath and Govindaraju (2001)) namely online HWR and offline HWR (refer figure 1.1).

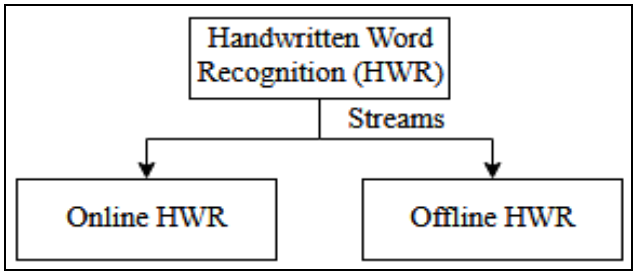


Figure 1.1: Streams of handwritten word recognition

1.2.1 Online Handwritten Word Recognition

In online HWR, one writes on an electronic surface such as an electronic tablet with a special pen and words or characters are recognized at real time as soon as it is written. Words or characters are captured as a sequence of strokes in these systems (Sen *et al.*, 2018). Online systems obtain the position of the pen as a function of time in the form of (x, y) coordinates, directly from the interface.

1.2.2 Offline Handwritten Word Recognition

It is the process of translating offline handwritten word into a format that is understood by machine. In offline HWR, the handwritten word is typically scanned from a paper document and made available in the form of a binary or gray scale image to the recognition algorithm. Offline HWR become more challenging due to shape of characters, great variation of character symbols and document quality (Bolan *et al.*, 2015). Therefore, offline word or character recognition is considered as a more challenging task than its online counterpart. Offline HWR is significantly different from online HWR as depicted in Table 1.1.

Table 1.1: Offline *Vs* online handwritten word recognition

Parameters	**Offline HWR**	**Online HWR**
Availability of strokes	No	Yes
Raw data requirement	Dots per inch	Samples per second
Writing media	Paper document	Digital pen on an electronic surface
Recognition speed	Lower	Sufficiently higher
Accuracy	Lower	Sufficiently higher

Researchers used several approaches/methodologies for the recognition of offline handwritten Devanagari words. Generally, these approaches/methodologies can be categorized into three categories as mentioned below (Kumar, 2016):

1.2.2.1 Segmentation Based or Analytical Approach

In this approach, the given word is segmented into individual components (or character parts or even character subparts) and each component (or character) is recognized and assigned a symbol and the resulting symbols are reconstructed (or reassembled) to know the identity of a word. Recognition is considered as a matching process which is usually done with the help of dictionary/library words (Garg *et al.*, (2015a); Shearme and Leach (1968)). The segmentation for obtaining each individual character is usually based on heuristics and done before the recognition of each character.

1.2.2.2 Segmentation Free or Holistic Approach

It is a word based approach, where a word is not broken down into individual components (or character parts). The word is treated as a single unit (or a whole) and there is no attempt to identify characters individually (Dasgupta *et al.*, 2016). Classifier is trained with a complete set of words by extracting their properties that helps to recognize the word as a whole. Challenges to this approach include complexity as the whole word is treated as a single indivisible entity and lower discrimination capabilities. The use of holistic approaches is usually restricted to areas where few word groups are used (small size vocabulary) and constrained to a fixed lexicon such as numbered words on cheques. These approaches have been less studied than part-based ones and recently have gained considerable interest.

1.2.2.3 Hybrid

In hybrid, the combination of segmentation based and holistic approaches is used.

1.3 APPLICATIONS AND CHALLENGES

Optical Character Recognition (OCR) has been a well-known area of research for last five decades (Panyam *et al.*, 2016). To identity the significance of OCR methods, in general, it is handy to present its applications and challenges.

1.3.1 Applications

Nowadays, there is a vast demand of techniques that consider automatic optical character or script identification or recognition. OCR systems can be designed for variety of applications such as (Chaudhuri and Pal (1997); Feng *et al.*, (2017)):

a) Postal service assistance by automatic sorting of postal documents.

b) Office and library automation.

c) Reading handwritten digits from bank cheques.

d) To help the visually handicapped when interfaced with a voice synthesizer.

e) Reading entrance examination forms.

f) Reading registration numbers from vehicle license plate images.

g) Processing of applications of victims and criminal records in police station.

h) For identification of type and recirculation of bottles in automats.

i) Text extraction and recognition of hand filled structured form documents.

j) Magnetic strip found on bank cards, membership cards, hotel door cards *etc.* contains information regarding say bank account number, name of card holder and expiry date of membership. This information can only be read by specially designed reader which sends data from card into computer for further processing.

k) The bar code is read optically using a visible red light, called as barcode scanner which reflects and transform the stored information on the barcode into digital information. It is necessary and helpful for retrieving information about the product such as product ID number, manufacturer and country of origin.

l) Magnetic Ink Character Recognition (MICR) system is mainly used for bank applications that can read information printed with a special ink containing iron particles.

m) Optical Mark Reading technology automatically reads location of marks made in pen or pencil.

A minor mistake in interpreting the words or characters can lead to error in the automation process such as wrong dispatch in postal service or wrong entry in entrance examination forms. Therefore, an accurate, robust and reliable handwritten character and word recognition by computer system would be greatly useful. The present work shall emphasize to develop a framework for handwritten Devanagari word recognition for postal automation.

Some major challenges need to be laid out and handled in order to achieve a successful automation. These challenges have been discussed in the following subsection.

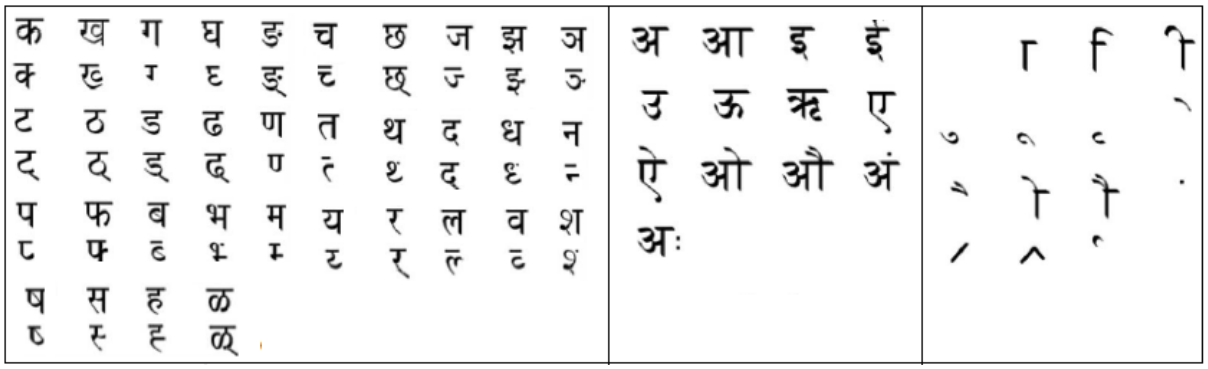
1.3.2 Challenges

Recognition of handwritten characters is a challenging task. Variations in writing styles from one person to another, as well as for a single individual from time to time, make this task harder (Sarkhel *et al.*, 2017). High quality or high resolution images (with some basic structural properties such as high differentiating text and background) are desired for achieving higher accuracy of OCR system. The recognition rate directly depends upon the quality of input image. In order to achieve a successful automation in OCR techniques, numerous errors have to be overcome which affect the quality of images dramatically. Error may arise due to aspect ratios, blurring and degradation, complex backgrounds, existence of uneven illumination, low resolution, multilingual content (Ye and Doermann, 2015), scene complexity, skewness, tilting or perspective distortion, variations of text layout or fonts and warping which will make it difficult for OCR systems to distinguish text from non-text (Liang *et al.*, (2008); Lopez-Antequera *et al.*, (2017); Mhiri *et al.*, (2018)).

In contrast with printed character or word recognition, HCR or HWR is a more challenging task due to many factors such as individual styles of inscription, speed of writing, size of letters, physical and mental state of the writer, overlap of letters, large symbolic set, character complexity and shape similarity. These factors may affect and cause problems in correct character recognition by computer system. Thus, there is a need of method(s) which overcomes these challenges or factors so that OCR systems produce correct machine-readable digital form, automatically.

1.4 OVERVIEW OF DEVANAGARI SCRIPT

Devanagari belongs to Brahmic family of scripts of India, Nepal, Tibet and South Asian subcontinent (Chaudhuri and Pal, 1997). Devanagari is the national language of India and generally spoken by 600 million people in India (Pratap and Arya, 2012). It is being used for writing numerous languages *viz.* Hindi, Sanskrit, Marathi, Konkani, Nepali including similar other languages of South Asian subcontinent (Pandey *et al.*, 2017). Devanagari script consists of 13 vowels (*Swar*), 34 consonants (*Vyanjan*) along with 14 modifiers of vowels (Ramteke and Rane, 2012) and *rakar* symbols. Figure 1.2 represents consonants and their corresponding half forms, vowels and modifiers of Devanagari Script (Jayadevan *et al.*, (2011b); Kumar *et al.*, (2018b)).



(a) **(b)** **(c)**

Figure 1.2: Devanagari script (a) Consonants and corresponding half forms (b) Vowels (c) Modifiers

Moreover, apart from above, it has compound or composite characters which may be formed by merging two or more basic characters. A vowel followed by consonant may take a modified shape, depending on whether the vowel is placed to the left, right, top or bottom of the consonant and are known as modifiers or *matras*. There is no idea of lower and upper characters in Devanagari script (Hindi language) unlike Roman script (English language). Furthermore, characters including text and digits are written from left to right in Devanagari script, unlike Urdu script. Devanagari script has its own specified composition rules for combining vowels, consonants and modifiers. An additional feature of Devanagari is existence of a horizontal line on the top of characters called header line or *shirorekha* (Bansal and Sinha, 2000). Word written in Devanagari script may be divided in strips viz. top, core and bottom. This header line divides the top and core strips whereas virtual base line divides the core and bottom strips. The knowledge of script is important in sense, if a person knows

script of a language then he/she can easily read the words pertaining to that scripts on the basis of his/her mental dictionary. Three strips of a word in Devanagari script are depicted in the following figure 1.3.

Figure 1.3: Three strips of a word in Devanagari script

2.1 LITERATURE REVIEW

There has been always a great need of research in the area of OCR for Indian languages even though there are many challenges (Elleuch *et al.*, 2016) and lack of commercial market (Jayadevan *et al.*, 2011b). Research towards Indian HCR/HWR has gained much attention in recent years even though first research on Devanagari recognition was reported by Sethi and Chatterjee (1977). Many methods for recognition of numeral/digit, word/text and isolated characters written in Devanagari script along with methods of script recognition/identification have been proposed in past few decades and are briefly outlined in the following section. It also includes the research work carried out by various researchers for Indian postal automation, signature verification, recognition of city names and automatic bank cheque processing/reading.

2.2 RELATED WORK

2.2.1 Numeral/Digit Recognition

Hanmandlu and Murthy (2007a) presented a handwritten recognition system for Hindi and English numerals based on fuzzy model. Authors achieved overall recognition rates of 95% and 98.4% for Hindi and English numerals respectively. Bhattacharya and Chaudhuri (2009) developed two databases for handwritten numerals of the two most popular Indian scripts namely Devanagari (22,556 samples) and Bangla (23,392 samples). Authors also, presented a multistage cascaded recognition scheme using wavelet-based multiresolution representations and Multi-Layer Perceptron (MLP) classifiers for recognition of mixed handwritten numerals (Devanagari, Bangla and English). Niu and Suen (2012) proposed a hybrid model for recognizing handwritten digits using CNN (Convolutional Neural Network) and SVM

(Support Vector Machine) as trainable feature and recognizer respectively. Authors used MNIST (Modified National Institute of Standards and Technology) digit database for automatic extraction of features from the raw images and generated the predictions. They achieved a recognition rates of 99.81% (without rejection) and 94.40% (with 5.60% rejection). Aggarwal *et al.* (2012) proposed an isolated handwritten Devanagari numerals recognition systems based on gradient feature and SVM classifier. Authors achieved 99.60% accuracy for standard dataset as provided by ISI (Indian Statistical Institute) Kolkata.

Arya *et al.* (2015) presented an offline Devanagari handwritten numeral recognition system using Gabor filter for feature extraction; nearest neighbor and SVM for classification taking database provided by ISI, Kolkata. In this study, authors tested for three filter sizes of 7×7, 19×19 and 31×31 to obtain optimal filter size and achieved the maximum recognition accuracy 98.06%. Furthermore, Dongre and Mankar (2015) presented a system for the recognition of Devanagari numerals and characters using structural and geometric features. Authors used MLP-Neural Network (MLP-NN) as classifier and experiment was conducted on 3000 handwritten samples of Devanagari numerals. They achieved 93.17% recognition accuracy using 40 hidden neurons for the same. Kumar *et al.* (2018a) presented a comprehensive survey on character and numeral recognition of non-Indic and Indic scripts. Moreover, authors presented major challenges/issues for character/numeral recognition in their work.

2.2.2 Word/Text Recognition

Ho *et al.* (1992) presented a method for word recognition based on an analysis of the shape of a word as a whole object. Madhvanath and Govindaraju (2001) studied the role of holistic paradigms in handwritten word recognition in their research work. Buse *et al.* (2002) proposed an offline word recognition system based on structural information. They extracted features using Gabor filter, classification was carried out using Fuzzy sets and obtained encouraging recognition rates. Liu and Shi (2003) proposed a format driven handwritten word recognition system based on segmentation and segmentation-free methods which can be used for many applications namely postal codes, titles or trademarks. Authors achieved very promising results in terms of segmentation and recognition rates on real data from US and UK postal zipcode. Bhowmik *et al.* (2008) developed a system for isolated handwritten Bangla words recognition, with a fixed Lexicon by means of a left-right Hidden Markov

Model (HMM). Authors used Genetic Algorithm (GA) to train the HMM and developed shape based direction encoding features. The proposed system was tested on the database consisting of 119 town names (West Bengal) each of which had 300 different handwritten samples. The training set and test set consist of 29,750 and 5950-word image samples respectively. To optimize the parameters of HMM, both non-discriminative and discriminative training procedures were applied iteratively. Kadhm and Hassan (2015) developed a handwriting word recognition system based on various feature extraction methods (namely connected components and zoning) and Support Vector Machine (SVM) classifier. Authors achieved recognition accuracy 96.317% on an Arabic handwritten words (AHDB database).

Dasgupta *et al.* (2016) presented a holistic offline handwritten cursive word recognition system using directional features based on Arnold transform followed by Hough transform. Cursive words written in English (CENPARMI database) were recognized using multiclass linear SVM and achieved overall accuracy of 87.19%. Furthermore, Kumar (2016) proposed a segmentation based approach for recognition of isolated hand printed Devanagari words (more than 3500 words were taken as database) and classification were carried out using Multi-Layer perceptron (MLP). Tamen *et al.* (2017) proposed an efficient multiple classifier system for Arabic handwritten words recognition. They proposed several combination rules between MLP, SVM and Extreme Learning Machine (ELM) classifiers trained with Chebyshev Moments (CM) and Statistical and Contour based Features (SCF). Authors achieved a global recognition rate equal to 96.82 % for IFN/ENIT database. Bhunia *et al.* (2018) proposed a novel method for cross-language handwritten text recognition and word spotting, considering three Indic scripts, namely Bangla, Devanagari and Gurumukhi. Furthermore, Sen *et al.* (2018) had proposed a novel Bangla word segmentation technique based on stroke-level busy zone formation procedure. Authors took 6500 online handwritten Bangla words as samples and achieved 98.45% correct segmentation accuracy.

2.2.3 Isolated Character Recognition

Arica and Yarman-Vural (2001) reviewed available offline handwritten character recognition methods along with its superiorities and weaknesses. Hanmandlu *et al.* (2007b) carried out the recognition of handwritten Hindi characters based on fuzzy models. Authors obtained overall recognition rate of 90.65% for a database of 4750 samples. Pal *et al.* (2008a)

developed a recognition system for offline handwritten Devanagari characters using combination of two classifiers namely MQDF (Modified Quadratic Discriminant Function) and SVM (Support Vector Machine) based on gradient and curvature features. Authors tested the proposed system on 36172 data samples and achieved accuracy 95.13%. Agrawal *et al.* (2009) proposed algorithm for identification of presence and position of vertical bar in the offline handwritten Hindi characters and achieved a coarse classification rate of 97.25%. Arora *et al.* (2010) discussed characteristics of the classification methods namely SVM and ANN (Artificial Neural Network) as applied for the recognition of handwritten Devanagari characters. Murthy and Hanmandlu (2011) proposed Zoning based Devanagari handwritten numeral and character recognition technique. Kubatur *et al.* (2012) proposed a neural network approach for recognition of Devanagari handwritten characters. Authors obtained recognition rates of up to 97.2% for a dataset of 2760 characters. Yadav *et al.* (2013) proposed an OCR system for printed Hindi text (Devanagari script) by means of Artificial Neural Network (ANN). In this work, three feature extraction techniques (histogram of projection based on mean distance, histogram of projection based on pixel value and vertical zero crossing) had been used to improve the rate of recognition. The testing was carried out at individual letter (1000 individual characters) level and paragraph (15 paragraphs consisting of 650 words) level and obtained recognition rates of 98.5% and 90% respectively.

Furthermore, Kale *et al.* (2013) in their work, achieved overall recognition rate of 98.25% and 98.36% for 3750 basic Devanagari and 11250 compound characters respectively considering Legendre moment as feature descriptor and Artificial Neural Network (ANN) as classifier. Dixit *et al.* (2014) proposed a system for recognition of handwritten Devanagari characters using wavelet based feature extraction and back propagation neural networks for classification. Experimentation was carried on 2000 samples and obtained accuracy of 70%. Furthermore, Jangid and Srivastava (2014) explored GLAC (Gradient Local Auto-Correlation) algorithm for recognition of Devanagari handwritten characters. Authors achieved recognition accuracies of 93.21% and 95.21% for ISIDCHAR (36172 character samples) and V2DMDCHAR (20305 character samples) datasets respectively. Acharya *et al.* (2015) introduced a new public image dataset of Devanagari handwritten characters (92,000 samples) and proposed a deep learning architecture for recognition of these characters. Authors achieved the highest test accuracy of 98.47% for given dataset using Deep Convolutional Neural Network (Deep CNN). Furthermore, Dongre and Mankar (2015)

presented a system for the recognition of Devanagari numerals and characters using structural and geometric features. Authors used MLP-Neural Network (MLP-NN) as classifier and experiment was conducted on 5375 handwritten samples of Devanagari characters. They achieved 82.7% recognition accuracy using 60 hidden neurons for the same. Shelke and Apte (2016) presented performance optimization and comparative analysis of neural networks for handwritten Devanagari character recognition. Authors implemented different neural networks and concluded that optimization at structural classification stage and feature extraction stage increases the recognition rate using database of 40,000 samples. Soora and Deshpande (2017) presented a comprehensive review of the feature extraction techniques for character recognition as feature extraction plays a major role in the performance of the character recognition.

2.2.4 Script Recognition

Ghosh *et al.* (2010) reviewed structure and visual appearance based methods of script identification developed by various researchers in detail. Rani *et al.* (2014) presented the effectiveness of Gabor filter banks with k-NN (k-Nearest Neighbors), SVM (Support Vector Machine) and PNN (Probabilistic Neural Network) classifiers to identify the scripts at line level from trilingual documents. Authors achieved recognition rate of 99.85% for trilingual documents (Gurmukhi, Hindi and English) using Gabor features with SVM classifier. Singh *et al.* (2015) in their research work discussed the advancements reported in the literature during the last few decades. Authors also discussed various feature extraction and classification techniques associated with the Offline Script Identification (OSI) of the Indic scripts. Shi *et al.* (2016) presented Discriminative Convolutional Neural Network (DiscCNN), a novel deep learning based method for script identification in natural images. The method combines deep features with discriminative mid-level representations. In addition, authors created a large scale dataset (namely SIW-13) containing 16,291 wild text images in 13 scripts for evaluation.

Obaidullah *et al.* (2017a) developed a word level document image dataset (39,000 words) of 13 different Indic languages (3,000 words per language) from 11 official scripts. Authors used five different classifiers namely Multi-Layer Perceptron (MLP), Fuzzy Unordered Rule Induction Algorithm (FURIA), Simple Logistic (SL), Library for Linear classifier (LibLINEAR) and Bayesian Network (BayesNet) for baseline results on script identification,

with three features namely Spatial Energy (SE), Wavelet Energy (WE) and the Radon Transform (RT), including their possible combinations. MLP provided better results when all features were used and achieved the bi-script accuracy of 99.24% (keeping Roman common), 98.38% (keeping Devanagari common) and tri-script accuracy of 98.19% (keeping both Devanagari and Roman common). Furthermore, Singh *et al.* (2017) presented a complete script recognition system for handwritten mixed-script documents. The document pages were first segmented into their corresponding text-lines and words. Then, the script recognition was done at word-level using texture-based features. The presented technique was applied on 100 mixed-script document pages written in Bangla or Devanagari text mixed with English words and achieved encouraging outcomes.

2.2.5 Applications

Roy *et al.* (2004) presented a system towards recognition of Bangla pincode numerals for Indian postal automation. Authors used structural features, combining Neural Network (NN) and tree classifier based approach for recognition of numerals and obtained overall accuracy 94.21%. Roy *et al.* (2005) presented a system towards Indian postal automation based on the recognition of pin-code and city name (English/Bangla) of the postal document based on water reservoir technique, NN and Non-Symmetric Half Plane-Hidden Markov Model (NSHP-HMM). Pal *et al.* (2008b) proposed a system for recognition of handwritten Bangla pin code string using water reservoir concept (for pre-segmentation), dynamic programming (for optimum segmentation), modified quadratic discriminant function (to compute the likelihood of a digit) and directional features. Authors achieved 99.08% reliability when rejection and error rates were 9.28% and 0.74%, respectively using their system on 2692 handwritten Bangla pin codes.

Pal *et al.* (2009a) proposed a scheme for recognition of Bangla handwritten city names using a lexicon-driven segmentation-recognition scheme for Indian postal automation. Authors tested their scheme on 84 different Bangla city names and achieved accuracy of 94.08%. Furthermore, Pal *et al.* (2009b) proposed a tri-lingual (English, Hindi and Bangla) 6-digit full pin-code string recognition. Authors obtained 99.01% reliability from their proposed system when error and rejection rates are 0.83% and 15.27%, respectively. Basu *et al.* (2010) proposed a system for recognition of the postal codes written in four scripts namely Latin, Devanagari, Bangla and Urdu. They implemented Hough transformation (to localize the

postal-code blocks), script independent unified pattern classifier (to classify the numeric postal codes), quad-tree based image partitioning technique (for effective feature extraction) and support vector machine in their work. Authors obtained recognition accuracies 95.55%, 95.63%, 97.15% and 96.20% for the Latin, Devanagari, Bangla and Urdu numerals respectively. Jayadevan *et al.* (2011a) discussed the important results reported in preprocessing, extraction, recognition and verification of handwritten fields on bank cheques and highlighted the positive directions of research. Pal *et al.* (2012) proposed a lexicon-driven method for multi-lingual (English, Hindi and Bangla) city name recognition for Indian postal automation. In the proposed scheme, a water reservoir concept (to pre-segment the slant corrected city names) and dynamic programming (to find optimum character segmentation) were implemented on 16132 Indian trilingual city names and achieved overall recognition 92.25%. Das *et al.* (2016) introduced a novel method to build up a multi-script off-line signature database aggregating many single script off-line databases, along with a statistical performance analysis method for a fair comparison between single and multi-script scenarios.

With advancements in digitization technology, OCR systems especially with Devanagari Script has gained an immense momentum over the last few decades. However, a lot of work still needs to be done for optimizations.

3.1 PHASES OF OCR SYSTEM

In this section, important phases (also stages or steps) of OCR system have been discussed. In general, these phases include image acquisition and pre-processing, segmentation, feature extraction and classification as depicted in Figure 3.1.

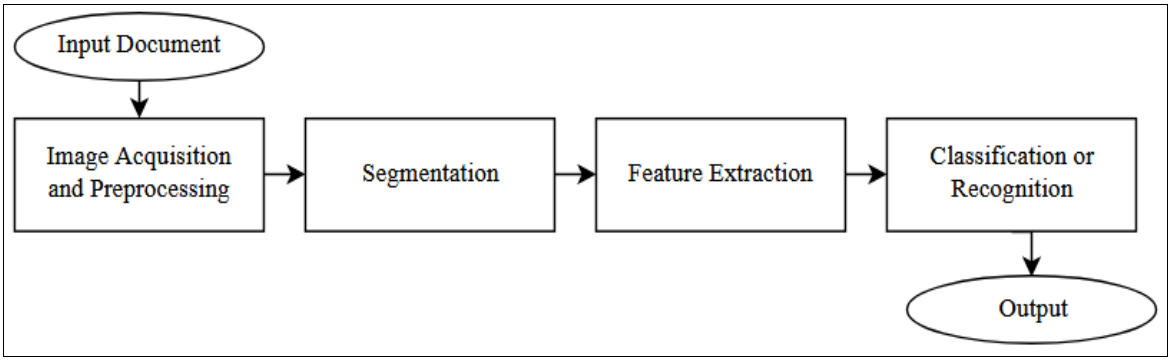


Figure 3.1: Phases of OCR system

Image acquisition and preprocessing: Handwritten paper based document is scanned to produce bitmap image or electronic form called digitization. It yields the digital image that can be fed to pre-processing phase. Pre-processing is a preliminary phase that aims to minimize the degradation in acquired image and produces a normalized bitmap image. Preprocessing generally include binarization, skeletonization, dilation of images, detection of edges, noise removal, thinning and filling, normalization (Pratap and Arya, 2012), perspective and page curl distortion (Ulges *et al.*, 2005), skew detection and correction (Chaudhuri and Pal, 1997) including the header line removal (Singh *et al.*, 2015).

Segmentation: In recognition, segmentation plays a significant role and it is used to break the scanned document into paragraphs, lines, words and characters (Malik, 2010). Segmentation of hand written characters is a challenging task due to a variety of writing styles (Hanmandlu *et al.*, 2009). Accuracy of OCR systems highly depends upon the detection of best segmentation points for paragraphs, lines, words and characters. Segmentation is divided into various parts namely line segmentation, word segmentation, zone segmentation and character segmentation (Garg *et al.*, (2011); Palakollu *et al.*, (2012); Ramteke and Rane (2012)).

Feature Extraction: Feature extraction plays a major role and important phase in pattern recognition (Cilia *et al.*, 2018). Features represent precise information extracted from segmented characters (symbols or words) that distinguish a particular character from other characters. Recognition accuracy of an OCR system also depends upon the selection of feature extraction technique. Feature extraction can be carried out in number of ways, however essential is to extract those features that can distinct dissimilar patterns or characters classes that exist. Features can be classified into two major categories namely statistical and structural.

- *Statistical features:* Characteristics of the distribution of pixel values on the bitmap image are captured as statistical features. Features can be calculated from statistical distributions of point's viz. moments, zoning, histograms or projection.

- *Structural features:* Structural features depict a pattern in terms of its topology and geometry by giving it local and global properties. These features are mainly

based on geometrical properties of a symbol or character viz. loops, directions of strokes, intersections of strokes and end points.

Many OCR researchers tried to amalgamate above two features *i.e.* statistical and structural together, in order to achieve better feature extraction.

Classification: Classification or recognition is a decision making phase that utilizes the features extracted in the previous phase for making class membership in the pattern recognition system (Singh *et al.*, 2015). It compares the input feature with stored pattern and results best suitable matching class for input. It can be carried out normally either using template or feature based methods.

- *Template-based method*: It involves the direct comparison between an unknown input pattern and original pattern. The amount of correlation between these two patterns is considered for classification or recognition.

- *Feature-based method:* It extracts features from the input pattern and uses these features for classification or recognition models *viz.* Artificial Neural Networks (ANN), Hidden Markov Models (HMM) (Bolan *et al.*, 2015), Support Vector Machines (SVM) (Elleuch *et al.*, (2016); Kumar (2009); Sahare and Dhok (2018)), Modified Quadratic Discriminant Function (MQDF).

4.1 RESEARCH ISSUES/GAPS

Although numerous research work on HWR in Roman, Japanese/Chinese and Arabic scripts exists but only limited work had been carried out towards offline HWR of Indian scripts specifically with non-isolated words (Roy *et al.*, 2016). Nowadays, postal automation machines for sorting post-letters are available in various countries viz. USA, UK, Canada, Japan, France and Germany but no machine is available for Indian Postal Automation (IPA) (Pal *et al.*, (2008b); Pal *et al.*, (2012)). It is a more challenging and hard to develop a system for IPA as India is a multi-script/multi-lingual country where 12 scripts (besides Roman script) namely Bangla, Devanagari, Gujarati, Gurumukhi, Kannada, Malayalam, Manipuri, Oriya, Sinhala, Tamil, Telugu and Urdu (Singh *et al.*, 2015) are used to write 22 official languages (besides English) namely Asamiya, Bangla, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam,

Manipuri, Marathi, Nepali, Odia, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu and Urdu (Basu *et al.*, (2010); Pal *et al.* (2009b)).

The recognition accuracy of OCR systems for IPA is promising, but more work need to be done. Moreover, standard OCR database (printed and handwritten) for all official Indian languages needs to be developed (Soora *et al.*, 2017) which would be made accessible publicly for research in Indic context. The available character segmentation approaches need to be improved so that these can handle a large variety of touching characters, which may occur fairly often in images obtained (Pandey *et al.*, 2017). Generally, people write postal address in two scripts say textual part in regional script and numeric part in Latin script.

Due to assorted writing styles of writers/people, sometimes unusual ligatures adjacent characters written in Devanagari frequently add misperception to recognize the words *viz.* "झ", "ड" and "इ". There is still a need of reseach in the field of Devanagri script for word, sentence and document recognition along with its semantics and lexicon (Pandey *et al.*, (2017); Pratap *et al.*, (2012)). This reseach work aims to provide a framework for offline handwritten Devanagri word recognition which will help to meet the growing needs for IPA, signature verification, city name recognition *etc.*

4.2 PROBLEM FORMULATION

Title of the present study is "**A Framework for Offline Handwritten Devanagari Word Recognition**".

The prime theme of this study is to analyze characteristics/features of the word, developing handwritten word recognition framework capable of identifying handwritten Devanagari words from image and converting it to editable text format for postal automation. Hindi (Devanagari script) is used by approximately 551.4 million people (according to 2001 census) in India (Singh *et al.*, 2015). Also, India's national language Hindi (Devanagari script) is world's third most popular language after Chinese and English (Jayadevan *et al.*, 2011b). As Hindi (Devanagari script) language has large number of character set, so it creates an open problem/challenge for researchers working in this area. Furthermore, handwritten Devanagari characters/words may suffer from distortion like other scripts (Roy *et al.*, 2016).

When any image containing Devanagari word is given as input, offline handwritten word recognition framework will perform certain processes and will convert it into equivalent form in a way to convert non-editable to editable form. There are an immense number of frameworks available for printed and handwritten character recognition for non-Indic scripts. But, only a limited number of systems are offered for character/numeral recognition of Indic scripts (Kumar *et al.*, 2018a). Therefore, there is a need to develop OCR system for offline handwritten Devanagari word recognition. For this, combination of various features and classifiers will also be explored to compensate the weakness of one approach by the strength of other approach.

4.3 CONTRIBUTION AND NEED OF THE PROPOSED RESEARCH

As OCR systems are capable of automatic translating of document images into equivalent character codes, so it helps in saving human energy as well as cost. OCR systems can be employed in various applications including postal addresses reading for automatic sorting, cheque verification, digitalization of ancient documents, text extraction and recognition from handwritten structured form documents, forensic and medical analysis, reading registration numbers from vehicle license plate images, automatic data entry *etc.* OCR systems viz. HWR can play a key role to improve or enhance the interaction between human and machine in many applications such as IPA. The handwritten text must be located, extracted, made free of artifacts stemming from the medium (underlines and background from the cheque leaf, boxes from forms, postal marks from the piece of mail), separated into lines if necessary and finally, into individual words before it can be recognized (Madhvanath *et al.*, 2001).

Although, the automatic recognition of printed text has achieved a great success rate, the current performance of HWR is not high. HWR has long been an active research area because of its vast potential applications (Roy *et al.*, 2016). So, finding potential need, inspiring challenges involved, huge application areas and interest committed to this field is the key motivation behind proposed research. At present there is no machine for IPA in India, hence there is need for developing more efficient methods of offline HWR with different scripts. Because of complex nature due to compound characters, modifiers, overlapping and touching (Garg *et al.*, 2015b) *etc.*, character segmentation and recognition is a tedious job in Indic scripts (*e.g.* Devanagari, Bangla, Gurumukhi, and

other similar scripts) (Roy *et al.*, 2016). HWR and spotting of low-resource scripts are difficult due to insufficient training data and it is costly for frequently collecting data of such scripts (Bhunia *et al.*, 2018). The present work will attempt to address the problem related to the interpretation of handwritten words in Devanagari script. Moreover, availability of sufficient and proper dataset is the key for success of OCR systems. Hence, in this research work approximately 10,000 samples of handwritten words/documents written in Devanagari script will be explored as a database/dataset.

5.1 RESEARCH OBJECTIVES

This thesis work shall broadly focus to develop a framework for offline handwritten Devanagari word recognition for postal automation. Based on the identified research gaps from the literature review some of the gaps are being considered for the research proposal. The broad objectives of the thesis work are outlined as follows:

I. To generate a corpus of handwritten Devanagari words for experimental work.

II. To explore existing features (structural and statistical) for offline handwritten Devanagari word recognition.

III. To propose and implement innovative features for offline handwritten Devanagari word recognition.

IV. To explore various classifiers for offline handwritten Devanagari word recognition.

V. To calculate, compare and analyze recognition accuracy by combining various features and classifiers.

In order to achieve these objectives, a detailed survey of literature on different stages of OCR system is being carried out. For this current study, more than 10,000 samples of handwritten Devanagari words will be explored as a dataset/database. This database/dataset will be used to evaluate the efficiency and effectiveness of the proposed word recognition framework.

5.2 AN OUTLINE OF THE RESEARCH PROPOSAL

The present work shall emphasize for the identification of handwritten Devanagari words as it might be helpful in many applications including postal automation.

5.2.1 Methodology

The proposed framework for offline handwritten Devanagari word recognition system will be implemented by considering the following steps/modules:

Step 1: More than 10,000 samples of handwritten Devanagari words will be explored as a database/dataset.

Step 2: Word-symbol segmentation/holistic approach will be carried out.

Step 3: In feature extraction phase, a few existing features will be explored and new innovative features will be proposed and implemented for handwritten Devanagari word recognition.

Step 4: Classification will be carried out using various classifiers.

Step 5: To calculate, compare and analyze recognition accuracy by combining various features and classifiers.

5.2.1.1 Assumptions

Following constraints will be considered for this work:

a) The handwritten words will be scanned at 300 dpi resolution.

b) Words will be considered noiseless and non-degraded.

c) No requirement of skew detection and correction.

d) Dataset/data will not contain any non-word items such as numbers, figures and tables *etc.*

5.2.2 Plan of Work

The plan of research work is summarized in the Table 5.1:

Table 5.1: Plan of work (Time-line chart)

Activity	Time Scale				
	Jan-Jun 2018	Aug-Dec 2018	Jan-Jun 2019	Aug-Dec 2019	Jan-Jun 2020
Literature Survey	■	■	■	■	■
Data Collection	■	■			
Preprocessing	■	■			
Feature Extraction and Selection		■	■		
Classification or Recognition			■	■	
Communication of Research Papers		■	■	■	
Documentation, Thesis Writing and Submission			■	■	■

Place of Work: Giani Zail Singh Campus College of Engineering & Technology (GZSCCET), Bathinda-151001 (Punjab), India.

Resource(s) Required: MATLAB Software *(Available at GZSCCET).*

Extent of External Assistance Required: YES *(Dr. Munish Kumar, Assistant Professor, Department of Computer Applications, MRSPTU, Bathinda as Co-Supervisor*)

BIBLIOGRAPHY

Acharya S, Pant AK, Gyawali PK (2015) Deep learing based large scale handwritten Devanagari character recognition. *Proceeding of Ninth International Conference on Software, Knowledge, Information Management and Applications* (SKIMA), 1-6.

Aggarwal A, Rani R, Dhir R (2012) Recognition of Devanagari handwritten numerals using Gradient features and SVM. *International Journal of Computer Applications*, 48(8):39-44.

Agrawal P, Hanmandlu M, Lall B (2009) Coarse classification of handwritten Hindi characters. *International Journal of Advanced Science and Technology*, 10:43-54.

Arica N and Yarman-Vural FT (2001) An overview of character recognition focused on off-line handwriting. *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews,* 31(2):216-233.

Arora S, Bhattacharjee D, Nasipuri M, Malik L, Kundu M, Basu DK (2010) Performance comparison of SVM and ANN for handwritten Devnagari character recognition. *International Journal of Computer Science Issues,* 7(6):18-26.

Arya S, Chhabra I, Lehal GS (2015) Recognition of Devnagari numerals using Gabor filter. *Indian Journal of Science and Technology,* 8(27):1-6.

Bansal V, Sinha RMK (2000) Integrating knowledge sources in Devanagari text recognition system. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans,* 30(4):500-505.

Basu S, Das N, Sarkar R, Kundu M, Nasipuri M, Basu DK (2010) A novel framework for automatic sorting of postal documents with multi-script address blocks. *Pattern Recognition,* 43:3507-3521.

Bhattacharya U, Chaudhuri BB (2009) Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 31(3):444-457.

Bhowmik TK, Parui SK, Roy U (2008) Discriminative HMM training with GA for handwritten word recognition. *IEEE Proceeding of Nineteenth International Conference on Pattern Recognition,* 1-4.

Bhunia AK, Roy PP, Mohta A, Pal U (2018) Cross-language framework for word recognition and spotting of Indic scripts. *Pattern Recognition,* 79:12-31.

Bolan S, Zhang X, Shijian L, Tan CL (2015) Segmented handwritten text recognition with recurrent neural network classifiers. *IEEE Proceeding of Thirteen International Conference on Document Analysis and Recognition* (ICDAR), 386-390.

Buse R, Liu ZQ, Bezdek J (2002) Word recognition using fuzzy logic. *IEEE Transactions on Fuzzy Systems,* 10 (1):65-76.

Chaudhuri BB and Pal U (1997) Skew angle detection of digitized Indian script documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 19(2):182-186.

Cilia ND, Stefano CD, Fontanella F, Freca ASD (2018) A ranking-based feature selection approach for handwritten character recognition. *Pattern Recognition Letters*. doi: 10.1016/j.patrec.2018.04.007

Das A, Ferrer MA, Pal U, Pal S, Diaz M, Blumenstein M (2016) Multi-script versus single-script scenarios in automatic off-line signature verification. *IET Biometrics,* 5(4):305-313.

Dasgupta J, Bhattacharya K, Chanda B (2016) A holistic approach for off-line handwritten cursive word recognition using directional feature based on Arnold transform. *Pattern Recognition Letters,* 79:73-79.

Dixit A, Navghane A, Dandawate Y (2014) Handwritten Devanagari character recognition using wavelet based feature extraction and classification scheme. *Proceeding of Annual IEEE India Conference* (INDICON), 1-4.

Dongre VJ, Mankar VH (2015) Devanagari offline handwritten numeral and character recognition using multiple features and neural network classifier. *IEEE Proceeding of Second International Conference on Computing for Sustainable Global Development* (INDIACom), 425-431.

Elleuch M, Maalej R, Kherallah M (2016) A new design based-SVM of the CNN classifier architecture with dropout for offline Arabic handwritten recognition. *Procedia Computer Science,* 80:1712-1723.

Feng Z, Yang Z, Jin L, Huang S, Sun J (2017) Robust shared feature learning for script and handwritten/machine-printed identification. *Pattern Recognition Letters,* 100:6-13.

Garg NK, Kaur L, Jindal MK (2011) The segmentation of half characters in handwritten Hindi text. *Proceeding of International Conference on Information Systems for Indian Languages* (ICISIL), 139: 48-53.

Garg NK, Kaur L, Jindal MK (2015a) Recognition of handwritten Hindi text using middle region of the words. *International Journal of Software Innovation* 3(4): 62-71.

Garg NK, Kaur L, Jindal MK (2015b) Segmentation of touching modifiers and consonants in middle region of handwritten Hindi text. *Pattern Recognition and Image Analysis* 25(3):413-417.

Ghosh D, Dube T, Shivaprasad AP (2010) Script Recognition-A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 32(12):2142-2161.

Hanmandlu M (2009) Segmentation of handwritten Hindi text: a structural approach. *International Journal of Computer Processing of Languages,* 22(1):1-20

Hanmandlu M, Murthy OVR (2007a) Fuzzy model based recognition of handwritten numerals. *Pattern Recognition,* 40:1840-1854.

Hanmandlu M, Murthy OVR, Madasu VK (2007b) Fuzzy model based recognition of handwritten Hindi characters. *IEEE Proceeding of Ninth Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications* (DICTA'07), 454-461.

Ho TK, Hull JJ, Srihari SN (1992) A word shape analysis approach to lexicon based word recognition. *Pattern Recognition Letters,* 13:821-826.

Jangid M, Srivastava S (2014) Gradient local auto-Correlation for handwritten Devanagari character recognition. *IEEE Proceeding of International Conference on High Performance Computing and Applications* (ICHPCA), 1-5.

Jayadevan R, Kolhe SR, Patil PM, Pal U (2011a) Automatic processing of handwritten bank cheque images: a survey. *International Journal on Document Analysis and Recognition*, 15(4):267-296.

Jayadevan R, Kolhe SR, Patil PM, Pal U (2011b) Offline recognition of Devanagari script: a survey. *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews,* 41(6):782-786.

Kadhm MS, Hassan AKA (2015) Handwriting word recognition based on SVM classifier. *International Journal of Advanced Computer Science and Applications,* 6(11):64-68.

Kale KV, Chavan SV, Kazi MM, Rode YS (2013) handwritten Devanagari compound character recognition using Legendre moment: an artificial neural network approach. *IEEE Proceeding of International Symposium on Computational and Business Intelligence* (ISCBI), 274-278.

Kubatur S, Sid-Ahmed M, Ahmadi M (2012) A neural network approach to online Devanagari handwritten character recognition. *IEEE International Conference on High Performance Computing & Simulation* (HPCS), 209-214.

Kumar M, Jindal MK, Sharma RK, Jindal SR (2018a) Character and numeral recognition for non-Indic and Indic scripts: a survey. *Artificial Intelligence Review*, 1-27. doi.org/10.1007/s10462-017-9607-x.

Kumar P, Saini R, Roy PP, Pal U (2018b) A lexicon-free approach for 3D handwriting recognition using classifier combination. *Pattern Recognition Letters,* 103:1-7.

Kumar S (2009) Performance comparison of features on Devanagari hand-printed dataset. *International Journal of Recent Trends in Engineering*, 1(2):33-37.

Kumar S (2016) A study for handwritten Devanagari word recognition. *IEEE International Conference on Communication and Signal Processing* (ICCSP), *1009-1014.*

Liang J, DeMenthon D, Doermann D (2008) Geometric rectification of camera-captured document images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 30(4):591-605.

Liu X and Shi Z (2003) A format-driven handwritten word recognition system. IEEE *Proceedings of the Seventh International Conference on Document Analysis and Recognition* (ICDAR'03), 1118-1122.

Liwicki M, Ebert S, Dengel A (2014) Bridging the gap between handwriting recognition and knowledge management. *Pattern Recognition Letters*, 35:204-213.

Lopez-Antequera M, Gomez-Ojeda R, Petko N, Gonzalez-Jimenez J (2017) Appearance-invariant place recognition by discriminatively training a convolutional neural network. *Pattern Recognition Letters*, 92:89-95.

Madhvanath S, Govindaraju V (2001) The role of holistic paradigms in handwritten word recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 23(2):149-164.

Malik L (2010) Recognition of handwritten Devanagari script. *International Journal of Pattern Recognition and Artificial Intelligence,* 24(5):809-822.

Mhiri M, Abuelwafa S, Desrosiers C, Cheriet M (2018) Hierarchical representation learning using spherical k-means for segmentation-free word spotting. *Pattern Recognition Letters,* 101:52-59.

Murthy OVR, Hanmandlu M (2011) Zoning based Devanagari character recognition. *International Journal of Computer Applications*, 27(4):21-25.

Niu XX, Suen CY (2012) A novel hybrid CNN-SVM classifier for recognizing handwritten digits. *Pattern Recognition,* 45:1318-1325.

Obaidullah SK, Santosh KC, Halder C, Das N, Roy K (2017a) Word-level multi-script Indic document image dataset and baseline results on script identification. *International Journal of Computer Vision and Image Processing,* 7(2):81-94.

Obaidullah SM, Santosh KC, Halder C, Das N, Roy K (2017b) Word-level thirteen official Indic languages database for script identification in multi-script documents. *Communications in Computer and Information Science,* 709:16-27.

Pal U, Chanda S, Wakabayashi T, Kimura F (2008a) Accuracy improvement of Devnagari character recognition combining SVM and MQDF. *Proceeding of Eleventh International Conference on Frontiers in Handwriting Recognition* (ICFHR), 367-372.

Pal U, Roy K, Kimura F (2008b) Bangla Handwritten Pin Code String Recognition for Indian Postal Automation. *Proceeding of Eleventh International Conference on Frontiers in Handwriting Recognition* (ICFHR), 290-295.

Pal U, Roy K, Kimura F (2009a) A lexicon-driven handwritten city-name recognition scheme for Indian postal automation. *IEICE Transactions on Information and Systems*, 92(5):1146-1158.

Pal U, Roy RK, Roy K, Kimura F (2009b) Indian multi-script full pin-code string recognition for postal automation. *IEEE Proceeding of Tenth International Conference on Document Analysis and Recognition* (ICDAR), 456-460.

Pal U, Roy RK, Kimura F (2012) Multi-lingual city name recognition for Indian postal automation. *IEEE Proceeding of International Conference on Frontiers in Handwriting Recognition* (ICFHR), 169-173.

Palakollu S, Dhir R, Rani R (2012) Handwritten Hindi Text Segmentation Techniques for Lines and Characters. *Proceedings of the World Congress on Engineering and Computer Science* (WCECS), 1:1-5.

Pandey RC, Dawadi BR, Sharma S, Basnet A (2017) Dictionary based Nepali word recognition using neural network. *International Journal of Scientific & Engineering Research*, 8(5): 473-479.

Panyam NS, Lakshmi VTR, Krishnan RK, Rao KNV (2016) Modeling of palm leaf character recognition system using transform based techniques. *Pattern Recognition Letters*, 84:29-34.

Pratap N, Arya S (2012) A review of Devnagari character recognition from past to future. *International Journal of Computer Science and Telecommunications*, 3(6): 77-82.

Ramteke AS, Rane ME (2012) Offline handwritten Devanagari script segmentation. *International Journal of Scientific & Technology* Research, 1(4):142-145.

Rani R, Dhir R, Lehal GS (2014) Gabor features based script identification of lines within a bilingual/trilingual document. *International Journal of Advanced Science and Technology*, 66:1-12.

Roy K, Pal U, Chaudhuri BB (2004) A system for joining and recognition of broken Bangla numerals for Indian postal automation. *Proceeding of the Fourth Indian*

Conference on Computer Vision, Graphics and Image Processing (ICVGIP), 641-646.

Roy K, Vajda S, Pal U, Chaudhuri BB, Belaid A (2005) A System for Indian postal automation. *IEEE Proceedings of the Eight International Conference on Document Analysis and Recognition (ICDAR'05),* 1060-1064.

Roy PP, Bhunia AK, Das A, Dey P, Pal U (2016) HMM-based Indic handwritten word recognition using zone segmentation. *Pattern Recognition,* 60:1057-1075.

Sahare P, Dhok SB (2018) Multilingual character segmentation and recognition schemes for Indian document images. *IEEE Access,* 6:10603-10617.

Sarkhel R, Das N, Das A, Kundu M, Nasipuri M (2017) A multi-scale deep quad tree based feature extraction method for the recognition of isolated handwritten characters of popular Indic scripts. *Pattern Recognition,* 71:78-93.

Sen S, Chowdhury S, Mitra M, Schwenker F, Sarkar R, Roy K (2018) A novel segmentation technique for online handwritten Bangla words. *Pattern Recognition Letters*, 1-8.

Sethi IK, Chatterjee B (1977) Machine recognition of constrained hand printed Devanagari. *Pattern Recognition,* 9(2):69-75.

Sharma MK, Dhaka VP (2016) Pixel plot and trace based segmentation method for bilingual handwritten scripts using feedforward neural network. *Neural Computing and Applications,* 27(7):1817-1829.

Shearme JN, Leach PF (1968) Some experiments with a simple word recognition system. *IEEE Transactions on Audio and Electroacustics,* 16(2):256-261.

Shelke S, Apte S (2016) Performance optimization and comparative analysis of neural networks for handwritten Devanagari character recognition. *IEEE Proceeding of International Conference on Signal and Information Processing* (IConSIP), 1-5.

Shi B, Bai X, Yao C (2016) Script identification in the wild via discriminative convolutional neural network. *Pattern Recognition,* 52:448-458.

Singh PK, Das S, Sarkar R, Nasipuri M (2017) Handwritten mixed-script recognition system: a comprehensive approach. *Advances in Intelligent Systems and Computing,* 515:787-795.

Singh PK, Sarkar R, Nasipuri M (2015) Offline script identification from multilingual Indic-script documents: a state-of-the-art. *Computer Science Review,* 15-16:1-28.

Soora NR, Deshpande PS (2017) Review of feature extraction techniques for character recognition. *IETE Journal of Research,* 1-16. doi:10.1080/03772063.2017. 1351323.

Tamen Z, Drias H, Boughaci D (2017) An efficient multiple classifier system for Arabic handwritten words recognition. *Pattern Recognition Letters,* 93:123-132.

Ulges A, Lampert CH, Breuel T M (2005) Document image dewarping using robust estimation of curled text lines. *IEEE Proceeding of Eighth International Conference on Document Analysis and Recognition* (ICDAR'05), 1001-1005.

Yadav D, Sanchez-Cuadrado S, Morato J (2013) Optical character recognition for Hindi language using a neural-network approach. *Journal of Information Processing Systems,* 9(1):117-140.

Ye Q, Doermann D (2015) Text detection and recognition in imagery: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 37(7):1480-1500.